Information Sciences 606 (2022) 669–687

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Key point-aware occlusion suppression and semantic alignment for occluded person re-identification

Shujuan Wang [a,b], Bochun Huang [a,b], Huafeng Li [a,b,*], Guanqiu Qi [c,*], Dapeng Tao [d], Zhengtao Yu [a,b]

[a] Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, Yunnan, PR China
[b] Key Laboratory of Artificial Intelligence in Yunnan Province, Kunming University of Science and Technology, Kunming 650500, Yunnan, PR China
[c] Computer Information Systems Department, State University of New York at Buffalo State, Buffalo, NY 14222, USA
[d] FIST LAB, School of Information Science and Engineering, Yunnan University, Kunming 650091, Yunnan, PR China

ARTICLE INFO

Article history:
Received 12 October 2021
Received in revised form 15 April 2022
Accepted 21 May 2022
Available online 28 May 2022

Keyword:
Person re-identification
Semantic alignment
Key point-aware
Occlusion suppression

ABSTRACT

Partial occlusion is a key factor affecting the performance of person re-identification (re-ID). Although some solutions have been specially designed for occluded person re-ID, the ambiguity of pedestrian appearances and complex backgrounds still pose great challenges. Therefore, a key-point-aware occlusion suppression and semantic alignment (POS) method is proposed in this study to alleviate the existing challenges in occluded person re-ID. This method consists of three main modules: key point-aware semantic alignment (KPA-SA), self-similarity guided feature discriminability enhancement (SGFDE), and global feature extraction under occlusion suppression (GFE-UOS). In particular, the proposed KPA-SA semantically aligns the activated areas corresponding to specific pedestrian key points (e.g., head, shoulders, legs, feet) in multiple channels of different images. In addition, according to the paired left and right pedestrian key points, a cross-fusion mechanism can be applied to information compensation to alleviate information loss in occluded areas. The proposed SGFDE utilizes the self-similarity of non-occluded information of the same pedestrian captured from different views to enhance the discriminability of pedestrian identity features and suppress interference information unrelated to pedestrian identity. The proposed GFE-UOS fuses the heat maps of different key points to suppress the negative impact of occlusion on global feature extraction. The comparative experimental results verify the effectiveness of the proposed POS and its superiority over state-of-the-art methods. The related source codes were released on https://github.com/huangdaichui/occ_reid.

1. Introduction

As an important technique in intelligent monitoring, person re-ID is used to determine whether pedestrian images collected by non-overlapping cameras contain the same pedestrian(s). Following the rapid development of deep learning, deep neural network-based person re-ID has progressed significantly in recent years [28,12,35,37,29,46,5,33]. Pedestrians are often occluded with varying degrees in real-world scenes; therefore, all the physical features of a pedestrian cannot be

* Corresponding author at: Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, Yunnan, PR China (H. Li); Computer Information Systems Department, State University of New York at Buffalo State, Buffalo, NY 14222, USA (G. Qi).
E-mail addresses: hfchina99@163.com (H. Li), qig@buffalostate.edu (G. Qi).

https://doi.org/10.1016/j.ins.2022.05.077
0020-0255/

presented by one image, which poses a great challenge to the existing person re-ID solutions. To solve this issue, occluded person re-IDs have drawn the attention of many researchers, and some feasible solutions have been proposed [4,27].

As shown in Fig. 1, owing to the differences caused by occlusion, camera views, and pedestrian postures, there is no correspondence between the same spatial positions in different images. If an unspecified feature encoder (FEN) is directly applied to this type of image, the discriminability of the extracted features is relatively low because of the non-correspondence of body parts at the same positions, thereby reducing the matching accuracy of pedestrian images. In addition, occlusion inevitably introduces irrelevant identity features to pedestrians, which further aggravates the ambiguity of pedestrian appearances. Although existing methods consider the poor recognition performance caused by the non-correspondence of feature semantics, they ignore the impact of occlusion on feature extraction [3,10,19,14,48]. Therefore, they often show unsatisfactory recognition performance in occluded person re-IDs.

Based on the a priori paired pedestrian key points (excluding the key point of the head), this study proposes a key **P**oint-aware **O**cclusion suppression and **S**emantic alignment (POS) method for occluded pedestrian re-ID. This method is primarily composed of three modules: key point-aware semantic alignment (KPA-SA), self-similarity guided feature discriminability enhancement (SGFDE), and global feature extraction under occlusion suppression (GFE-UOS).

In contrast to the existing key point-based person re-ID methods [4,27] and feature semantic alignment methods [10,3], the proposed module KPA-SA uses a pedestrian key point-aware method to align the local pedestrian areas concerned by different feature channels and enrich the information of the corresponding local areas. Owing to the differences caused by occlusion and views, the features at the same spatial positions of different images do not correspond to each other. The proposed FEN can activate the areas corresponding to the same pedestrian key points on the feature channels of the same serial numbers in different images. As shown in Fig. 2, once the proposed network can activate and focus on key point-aware areas, the semantic alignment of image features from different pedestrian images can be achieved. In the above process, the heat maps generated by the key-point detection model enable the FEN to achieve the semantic alignment of key points under the constraints of key-point categories. Consequently, the above non-correspondence issue of the features can be solved. Moreover, the symmetric information compensation of pedestrian key points is explored using a cross-fusion mechanism to alleviate the information loss of occluded areas. Since the key points of a pedestrian are distributed in different areas of the pedestrian body, KPA-SA can extract the features of the corresponding non-occluded area of each key point driven by key point perception. Therefore, when the area corresponding to a certain key point is occluded, the proposed method can extract the discriminative features from the areas corresponding to the other key points. Unlike KPA-SA, most traditional person re-ID models focus only on the most discriminative areas of pedestrians. Once the focused area is occluded and invisible, the performance of the corresponding model is compromised. This issue is effectively solved by KPA-SA.

The proposed SGFDE module enhances the feature discriminability of non-occluded areas. The differences in the occluded areas of the same pedestrians from different views and the feature similarities of the non-occluded areas are comprehensively analyzed. Therefore, the discriminability of pedestrian identity features is enhanced, and the occlusion suppression is realized simultaneously. The proposed GFE-UOS module first fuses the heat maps of different key points to form a global heat map. The obtained global heat map is then used to assist the extraction of global features, in which global average pooling (GAP) and global maximum pooling (GMP) are integrated to strengthen the discriminative features of pedestrians. This module effectively not only extracts the discriminative features of pedestrians from non-occluded areas, but also suppresses occlusion and complex background simultaneously. The proposed POS is applied to six challenging benchmarks. The comparative experimental results verify that the proposed POS achieves better overall performance than the state-of-the-art methods for occluded person re-ID.

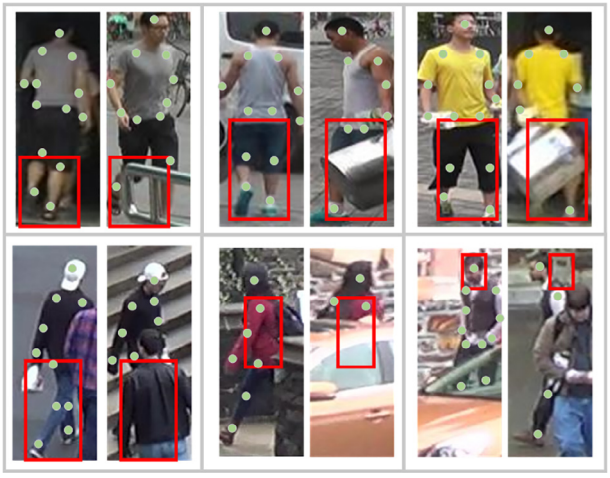

Fig. 1. Misalignment examples of different local areas of pedestrian images captured from different views. Due to various factors such as occlusion and different viewing angles, the key points of different pedestrian images marked in sage green dots are distributed in different spatial positions of different images. So, the same spatial positions of different images marked in red frames correspond to the features of different areas.

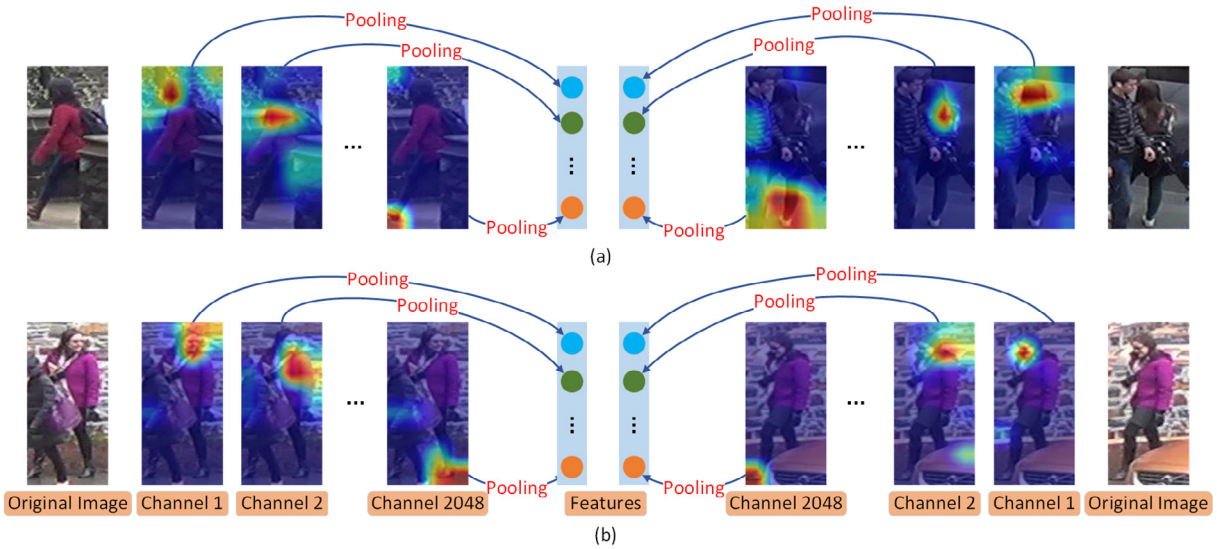

Fig. 2. Illustration of key point-aware semantic alignment in the KPA-SA. The responses of pedestrian key points (e.g. head, shoulders, legs, feet, etc.) shown in heat maps are focused by the proposed FEN. Red denotes the strong activation in heat maps. The first and last columns present the original images. The feature maps of each original image has 2,048 channels. The activated areas corresponding to the same pedestrian key points from two different images are aligned on the channel with the same serial number. (a) shows the alignment of the activated areas corresponding to the same pedestrian key points from two images of different pedestrians. (b) shows the alignment of the activated areas corresponding to the same pedestrian key points from two images of the same pedestrian.

Three main contributions of this paper are summarized as follows.

- The KPA-SA module is developed to enable specific feature channels focus only on the areas corresponding to the key points. So, the semantic alignment of features from non-occluded pedestrian areas and the discriminability enhancement of local area features are realized. Moreover, according to the paired left and right pedestrian key points, an information compensation solution for the left and right key points is proposed to alleviate the information loss in occluded areas.
- An SGFDE module is proposed to further explore the non-occluded areas. The diversity in the occluded areas of pedestrian images captured from different views and the self-similarity of pedestrian identity features are utilized to facilitate occlusion suppression and discriminability enhancement of pedestrian identity features.
- A GFE-UOS module is proposed to achieve the extraction of non-occluded global pedestrian features by the comprehensive utilization of the generated pedestrian key-point heat maps. The complementarity of both global and salient pedestrian features is used to enhance the discriminability of the pedestrian identity features. According to the fused key-point heat map, the occluded areas of pedestrians and complex backgrounds are effectively suppressed, and the pedestrian identity features draw great attention from the FEN.

The rest of this paper is organized as follows: Section 2 discusses the related work; Section 3 specifies the proposed solution in detail; Section 4 presents the comparative experiments and analyzes the corresponding experimental results, and Section 5 concludes this paper.

2. Related Work

The re-ID of occluded pedestrians can be categorized into two types, occluded person re-IDs and partial person re-IDs. Occluded person re-ID focuses on the re-ID of partially invisible pedestrians caused by occlusions. Partial person re-ID mainly involves the re-ID of incomplete pedestrian images caused by occlusions or misdetections. There is no obvious difference between the above two types of person re-IDs. Occluded person re-ID methods are often suitable for scenes with partial person re-ID. However, partial person re-ID methods often consider incomplete detection of the human body, and the detected results do not contain any occlusions. Therefore, it is difficult to apply partial person re-ID methods to the re-IDs of occluded pedestrians.

2.1. Occluded Person Re-ID

Pedestrian occlusions are unavoidable in the real-world environments. Therefore, person re-ID related applications are necessary to consider the impact of occluded pedestrians on the recognition performance. Zhuo et al. [49] used occluded/non-occluded binary classification loss to determine whether samples ware from an occluded or full-body sample set. In

a subsequent work, Zhou *et al.* [50] designed a teacher-student learning framework to extract the robust features of occlusion and highlight the discriminative areas using the predicted saliency maps. He *et al.* [7] proposed a re-ID method for occluded pedestrians without alignment. Both the fully convolutional network and pyramid pooling were first used to extract spatial pyramid features. Subsequently, a foreground-aware pyramid reconstruction matching method without alignment was proposed to calculate the matching scores among the occluded pedestrians.

In a recent study, Gao et al. [4] proposed a pose-guided visible part matching method that can jointly learn the discriminative features with pose-guided attention and self-mine partial visibility in an end-to-end framework. Zhang *et al.* [36] used the relationship between person re-ID and semantic segmentation to construct a person re-ID network that is robust to semantic perception and occlusion. Semantic segmentation is mainly used to filter out the impact of occlusion on pedestrian feature extraction. Tan *et al.* [25] proposed a multi-head self-attention network to remove unimportant information from pedestrian images and obtain key local information for occluded person re-ID. Pedestrian key points can play a supporting role in improving the discriminability of pedestrian features [17,27]. Miao *et al.* [17] used the detection of key points to estimate pedestrian postures; thus, non-occluded pedestrian features can be extracted from occluded pedestrian images. Wang *et al.* [27] used the graph relationship between the key points of the human body to extract discriminative features from occluded pedestrian images. In contrast to the above methods, the proposed solution uses pedestrian key points as a guide to promote the FEN to activate the same key point areas on the feature channels of the same serial numbers in different images to achieve semantic alignment for the improvement of feature discriminability. Additionally, both information loss and interference caused by occlusion can be effectively reduced in the extraction of discriminative features using the proposed cross-fusion mechanism and self-similarity calculation of identity features.

2.2. Partial Person Re-ID

Owing to the incomplete detection and limitation of camera views, only partial areas containing pedestrians are shown in one captured image, when pedestrians are occluded. Similar to occluded person re-ID, partial person re-ID [41] aims to match partial query images with holistic images in a gallery set. Zheng et al.[41] proposed a global-to-local matching model to obtain spatial layout information. He *et al.*[6] reconstructed the feature maps of partial queries from holistic pedestrian images and further avoided the impact of a cluttered background using fore-background heat maps. Sun *et al.*[23] proposed a visibility-aware part model that obtains the visibility of perceptive areas by self-supervised learning. To match a pair of pedestrian images of different sizes, deep spatial feature reconstruction was proposed to avoid explicit alignment. Luo *et al.*[16] proposed a deep partial re-ID framework based on pairwise spatial transformer networks, which can be obtained by training the existing holistic person datasets. Both the inconsistency of partial pedestrian images and incomplete pedestrian features limit the further improvement of the aforementioned methods. In contrast to these methods, the proposed POS makes full use of the paired left and right pedestrian key points and key point-aware semantic alignment; thus, it can effectively solve the inconsistency of partial pedestrian image features and further improve the performance of partial person re-ID.

3. Methodology

3.1. Overview

Given a training set $D_t = \{x_i, y_i\}_{i=1}^{N_t}$, where N_t is the total number of images in D_t, x_i is the i-th image, $y_i \in \{1, 2, \ldots, L_t\}_{i=1}^{N_t}$ is its corresponding identity label, and L_t is the total number of pedestrians. Suppose that the feature maps $\bar{\boldsymbol{F}}_i = \{\boldsymbol{E}(\boldsymbol{x_i})\}$ are obtained after the sample x_i is input to the backbone. As shown in Fig. 3, the proposed POS is composed of three modules: KPA-SA, SGFDE, and GFE-UOS. The KPA-SA module is used to realize the semantic alignment of features. The SGFDE module is applied to enhance the discriminability of non-occluded information. The GFE-UOS module is used to achieve the extraction of non-occluded global features for pedestrian identity matching. As shown in Fig. 3, since x_j only assists x_i to suppress the occluded information, the input image x_j as the only input of SGFDE highlights the discriminability of the non-occluded information in x_i.

After the feature map $\bar{\boldsymbol{F}}_i$ is input into the KPA-SA module, the heat maps corresponding to the key points are generated by the existing detection model of pedestrian key points [22]. In this paper, key points include head, left and right shoulders, left and right elbows, left and right hands, left and right crotches, left and right knees, and left and right ankles. The KPA-SA module uses the features extracted by heat maps to build the key point-aware capability under the action of key-point classifiers. Moreover, a cross fusion mechanism is embedded in the KPA-SA module for the compensation of occluded symmetric information. In the SGFDE module, the dissimilarity of occlusion across multiple images of the same pedestrian captured from different views is used to achieve the occlusion suppression and discriminability enhancement of pedestrian identity features simultaneously. In the GFE-UOS module, the input feature maps $\bar{\boldsymbol{F}}_i$ and the fused heat maps of key points are multiplied element by element to achieve occlusion suppression and promote the extraction of global non-occluded features. The proposed solution assumes that a pedestrian image has $K + 1$ categories of key points (K categories are from the paired key points and the remaining one category is from the human head) and a total of M key points.

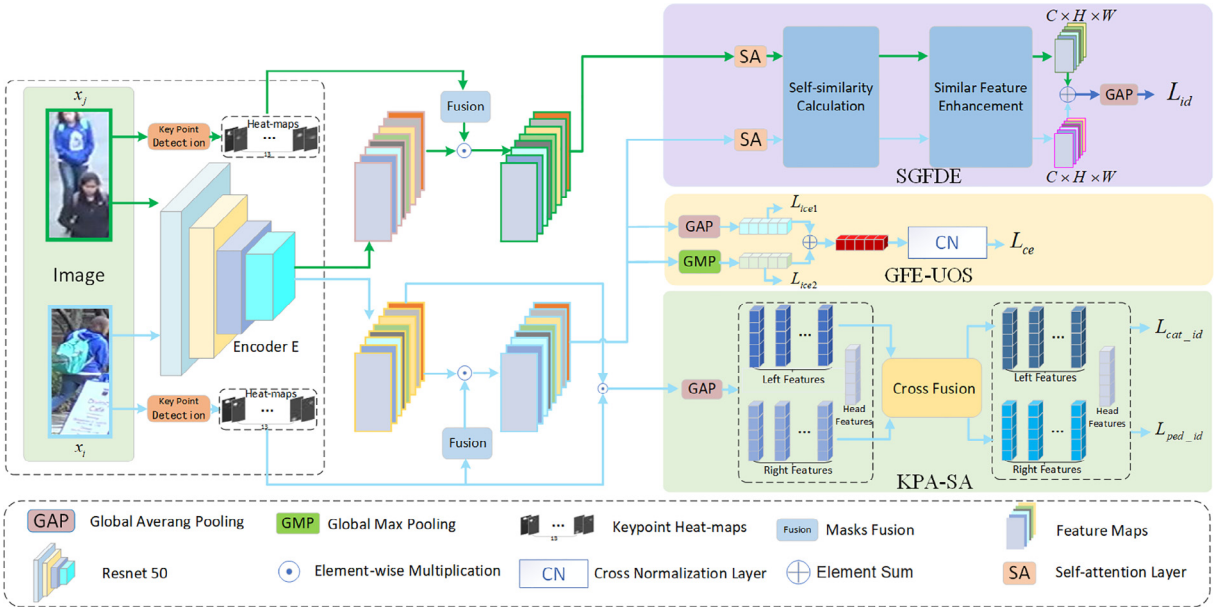

Fig. 3. Overall workflow of the proposed POS method. It consists of KPA-SA, SGFDE, and GFE-UOS three modules. Images x_j and x_i show the same pedestrian from different views, and the image x_j is used to assist the input image x_i to extract the discriminative pedestrian identity features. In POS, the key-point heat maps are first obtained after processing the input images by the encoder E and key-point detection model, and then the obtained key-point heat maps are fused to generate the global feature heat map. The obtained global heat maps and key-point heat map are used to obtain the global and local features corresponding to key points. The feature discriminability enhancement of the obtained results is achieved using the subsequent KPA-SA, SGFDE, and GFE-UOS modules.

3.2. Key Point-aware Semantic Alignment

Owing to the differences in occlusion, posture, and camera views, spatial misalignment as one key factor affecting feature discriminability often appears in pedestrian images. Although some existing methods can alleviate this issue, they do not consider the impact of occlusions on recognition performance [3,10]. According to existing research, the activated local responses of different channels correspond to the specific body parts [3,10]. Based on this principle, a KPA-SA module is proposed to alleviate the impact of both image space misalignment and occlusions on pedestrian identity matching. Specifically, the key-point heat maps are used to obtain local features as follows:

$$\boldsymbol{f}_l^k = GAP(M_l^k \odot \boldsymbol{E}(x_i)), \tag{1}$$

$$\boldsymbol{f}_r^k = GAP(M_r^k \odot \boldsymbol{E}(x_i)), \tag{2}$$

$$\boldsymbol{f}_h = GAP(M_h \odot \boldsymbol{E}(x_i)), \tag{3}$$

where $k \in [1, 2, \cdots, K], i \in [1, 2, \cdots, N_t], f_l^k, f_r^k$, and f_h denote the local features corresponding to the left and right key points of the k-th category of key points (except the head) and the key point of head, respectively. M_l^k and M_r^k represent the left and right key-point heat maps of the k-th category of key points (except the head), respectively. M_h represents the blue heat map corresponding to head. $\odot$means the element-wise multiplication. GAP is short for global average pooling. In Eqs. (1)–(3), the heat maps M_l^k, M_l^k and M_h are generated by the pre-trained HR-Net [22]. A high-resolution subnetwork is used as the first stage of HR-Net, and then high-to-low resolution subnetworks are gradually added one by one to compose multiple stages. Additionally, multi-resolution subnetworks are connected in parallel. Multi-scale fusion is repeated by exchanging information in parallel multi-resolution subnetworks throughout the process. The key points are estimated on the high-resolution representations output by HR-Net.

After the pedestrian image is input into HR-Net, the network estimates the position of each pedestrian key point and calculates the probabilities of key points at each pixel. The obtained probabilities are used as the pixel values at the corresponding pixels in the heat maps (refer to [22] for the related details). Since the pedestrian's body exhibits left–right symmetry, except for the key points involving head as the centerline, the remaining key points (consisting of left and right key points) are in pairs. If partial local areas of pedestrians are occluded, according to the paired left and right of pedestrian key points, the integration of symmetric key-point features is achieved using the following cross-fusion mechanism:

$$\tilde{\boldsymbol{f}}_l^k = \boldsymbol{w}_l \odot \boldsymbol{f}_l^k + (\boldsymbol{1} - \boldsymbol{w}_l) \odot \boldsymbol{f}_r^k, \tag{4}$$

$$\tilde{\boldsymbol{f}}_r^k = \boldsymbol{w}_r \odot \boldsymbol{f}_r^k + (\boldsymbol{1} - \boldsymbol{w}_r) \odot \boldsymbol{f}_l^k, \tag{5}$$

where $\boldsymbol{w}_l$ and $\boldsymbol{w}_r$ as the learnable weight vectors control the left and right symmetric key points, respectively. $\boldsymbol{1}$ is a vector, and the elements of $\boldsymbol{1}$ are 1. After integration, $\tilde{\boldsymbol{f}}_l^k$ and $\tilde{\boldsymbol{f}}_r^k$ represent the area features corresponding to the left and right key points, respectively. The detailed process of the proposed cross-fusion is shown in Fig. 4.

In the above process, when the corresponding area of a key point is occluded, a certain compensation of the occluded area can be achieved by taking advantage of the symmetry of the pedestrian's body. So, the information loss in the occluded side of the pedestrian can be effectively alleviated. Each element in the feature vector of each pedestrian image is obtained from the feature map on the corresponding feature channel by GAP or GMP. Once the encoder $\boldsymbol{E}$ activates the local area responses corresponding to the same pedestrian key points on the feature channels of the same serial numbers in different images, the cross-image semantic alignment can be realized. So, the constraints on the corresponding areas of key points of each human body after integration are shown as follows:

$$\boldsymbol{L}_{cat\_id} = -\sum_{i=1}^{N}\sum_{m=1}^{M}\delta_i^m \log(\boldsymbol{W}(\tilde{\boldsymbol{f}}_i^m)), \tag{6}$$

$$\boldsymbol{L}_{ped\_id} = -\sum_{i=1}^{N}\sum_{m=1}^{M}\delta_i \log(\boldsymbol{W}_m(\tilde{\boldsymbol{f}}_i^m)), \tag{7}$$

where $\tilde{\boldsymbol{f}}_i^m \in \{\tilde{\boldsymbol{f}}_r^1, \tilde{\boldsymbol{f}}_l^1, \tilde{\boldsymbol{f}}_r^2, \tilde{\boldsymbol{f}}_l^2, \cdots, \tilde{\boldsymbol{f}}_r^K, \tilde{\boldsymbol{f}}_l^K, \boldsymbol{f}_h\}$ is the feature vector corresponding to the m-th key point of the i-th image x_i by Eq. (4) or (5). N is the batch size. $\boldsymbol{W}$ is the classifier of key-point categories. $\boldsymbol{W}_m$ is the pedestrian identity classifier corresponding to $\tilde{\boldsymbol{f}}_i^m$. δ_i^m is the category label corresponding to the m-th key point of the i-th image x_i. δ_i is the identity label of $\tilde{\boldsymbol{f}}_i^m$, and the value of δ_i equals 1 only at y_i. The loss function $\boldsymbol{L}_{cat\_id}$ shown in Eq. (6) enables the FEN to extract pedestrian features related to key-point categories from the corresponding areas of key points; thus, the key point-aware semantic alignment is realized. The loss function $\boldsymbol{L}_{ped\_id}$ shown in Eq. (7) enables the FEN to extract the features related to pedestrian identities from the corresponding areas of key points. So, the feature discriminability is enhanced. In the loss function $\boldsymbol{L}_{cat\_id}$, $\boldsymbol{W}$ is used to assign $\tilde{\boldsymbol{f}}_i^m$ to the category corresponding to the m-th key point of the i-th image x_i, which can promote the FEN to achieve semantic alignment. In fact, the FEN focuses only on specific local areas of pedestrians on different feature channels. The obtained result by performing GAP on a single feature channel also corresponds to the areas focused by the FEN. As shown in Fig. 5, the classifier $\boldsymbol{W}$ with a fully connected layer is first used to assign the corresponding weights to the output results of cross fusion after GAP, and then the weighted sum result is classified into the corresponding key-point category.

The fully connected layer of the classifier $\boldsymbol{W}$ has the same weight in different images. To achieve the correct classification, the FEN needs to focus on the areas of the same key-point category on the feature channels, and align these feature channels with the same channel number. Therefore, the semantic alignment can be achieved. At the same time, the channel alignment also establishes the relationship between key points and the corresponding feature channels, which promotes the FEN to focus on the relevant areas of key points.

In addition, the pedestrian key points are distributed in all key body parts of pedestrians. The minimization of the loss functions $\boldsymbol{L}_{cat\_id}$ and $\boldsymbol{L}_{ped\_id}$ enables the FEN to extract the features of the areas corresponding to all non-occluded key points under the guidance of key point-aware constraints. So, the extracted features are complete and reliable. The proposed KPA-SA module is conducive to enhancing the discriminability of features. As shown in Fig. 6(b), under the constraints of key-point categories, the encoder $\boldsymbol{E}$ effectively activates local responses corresponding to key points. Since the encoder $\boldsymbol{E}$ focuses on the relevant areas, it can help the FEN focus on the relevant areas of key points and learn more reliable and complete

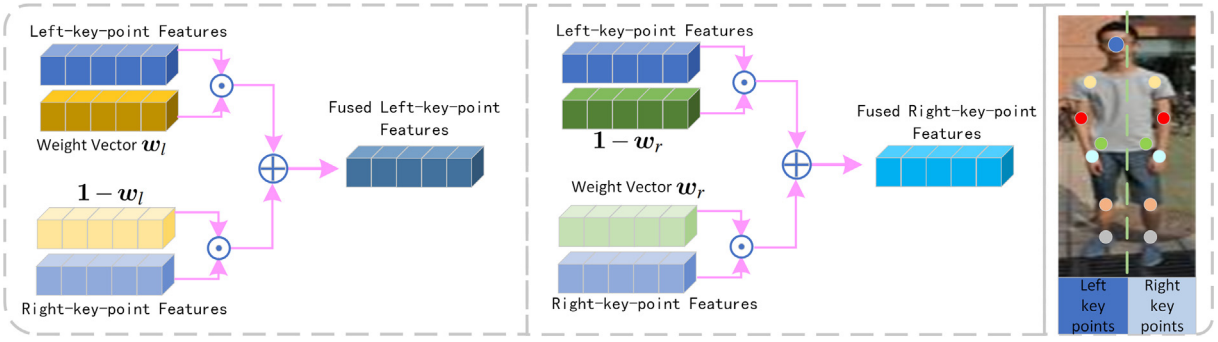

Fig. 4. Cross-fusion of the corresponding areas of the paired pedestrian key points (excluding the key point of the head). The left half shows the process of obtaining the fused left-key-point features by the left-key-point feature vector under the action of the weights w_l and 1-w_l. The right half shows the process of obtaining the fused right-key-point features by the right-key-point feature vectors under the action of the weights w_r and 1-w_r.

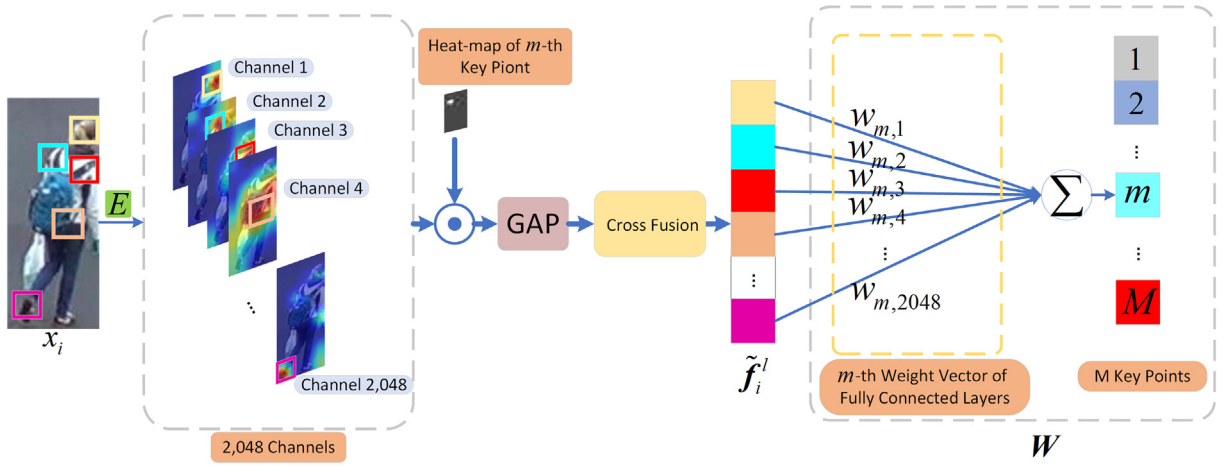

Fig. 5. Implementation process of the key point-aware semantic alignment. $w_{l,h}$ ($l \in \{1,2,\cdots,M\}, h \in \{1,2,\cdots,2,048\}$) denotes the weight of the fully connected layer on the h-th channel corresponding to the m-th key point. The pedestrian body parts marked in different color frames correspond to the heat maps (the activated local responses) marked in the same corresponding color frames on different channels.

pedestrian features rather than extracting specific local pedestrian features. Therefore, the proposed person re-ID model is robust to both partial occlusion and change of camera view.

Fig. 6(c) and (d) show the changes in the activation areas of one channel before and after adding the proposed KPA-SA module. After adding the proposed KPA-SA module, the encoder ***E*** can activate the same spatial areas of the corresponding key points on the feature channels with the same channel number across multiple different images. Therefore, the encoder ***E*** can achieve semantic alignment across multiple images. When the encoder ***E*** lacks the constraints of the proposed KPA-SA module, the activated areas of the same feature channels of different images are not consistent as shown in the first column of Fig. 6(c) and (d). Without the constraints of the proposed KPA-SA module, the encoder ***E*** cannot effectively activate the spatial local responses corresponding to the key points on the same channel across multiple images. So, the semantic alignment cannot be achieved by the encode ***E***.

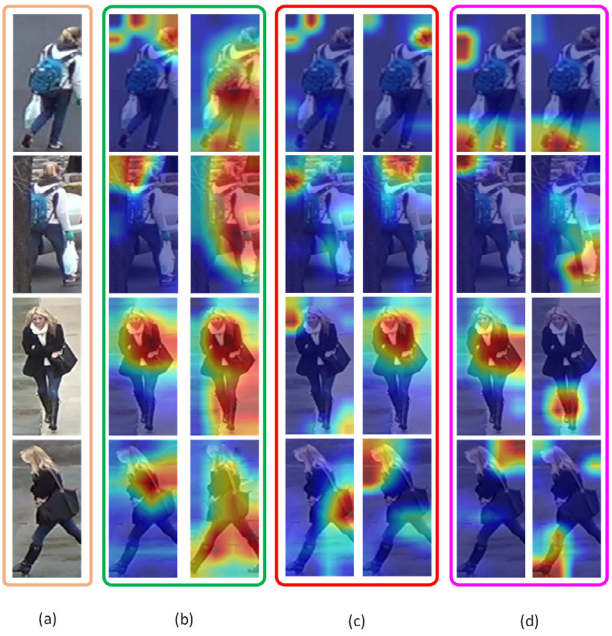

Fig. 6. Illustration of the effectiveness of the proposed KPA-SA module. Column (a) shows source images. The first column in (b) shows the areas focused by the encoder E without the KPA-SA module, and the second column in (b) shows the areas focused by the encoder E after adding the KPA-SA module. The first column of (c) and (d) shows the areas activated by the encoder E on a specific channel without the KPA-SA module, and the second column of (c) and (d) shows the areas activated by the encoder E on a specific channel after adding the KPA-SA module.

3.3. Self-similarity Guided Discriminability Enhancement

The pedestrian images used in person re-ID are captured by non-overlapping cameras. Under different views, the occlusions are often different, but the identity features of the same pedestrian usually remain same [39]. According to the self-similarity of pedestrians and the differences of both occlusion and background under different views, the SGFDE module is proposed to improve the feature discriminability of non-occluded areas. As shown in Fig. 7, the SGFDE includes two parts: self-similarity calculation and similar feature enhancement. Before the self-similarity calculation, the feature maps $\bar{\boldsymbol{F}}_i$ and $\bar{\boldsymbol{F}}_j$ of x_i and x_j are first filtered by the corresponding fused key-point heat maps $\boldsymbol{M}_{f,n} = \boldsymbol{M}_{h,n} + \sum_{k=1}^{K}(\boldsymbol{M}_{l,n}^k + \boldsymbol{M}_{r,n}^k)$, where $n = i$ or $j, M_{l,n}^k, M_{h,n}$, and $M_{r,n}^k$ correspond to M_l^k, M_h and M_r^k, respectively, and denote the corresponding key-point heat maps of image x_n.

Next, the filtered feature maps are input to the self-attention (SA) layer [30] to generate the input feature maps $\boldsymbol{F}_{i,s} = \mathrm{SA}(\boldsymbol{M}_{f,i} \odot \bar{\boldsymbol{F}}_i)$ and $\boldsymbol{F}_{j,s} = \mathrm{SA}(\boldsymbol{M}_{f,j} \odot \bar{\boldsymbol{F}}_j)$ of the self-similarity calculation, and a 1×1 convolution is used to fuse the obtained results on the feature channels. After that, the reshape operation is performed. Finally, the result of x_j obtained in the above process is multiplied by the transposed result of x_i to obtain a similarity matrix. As shown in Fig. 7, the similar feature enhancement marked in a red dashed line frame has two branches. In one branch (flowchart is marked in light blue), the result of self-similarity calculation is sequentially processed by softmax, reshape, and GMP on the feature channels, and then the obtained result and the feature map of x_i are multiplied element by element. In another branch (flowchart is marked in green), the result of self-similarity calculation is sequentially processed by softmax, transpose, reshape, and GMP on the feature channels, and the obtained result and the feature map of x_j are then multiplied element by element. For image x_i, the above process can be formulated as follows:

$$\boldsymbol{F}_i = GMP(Res(Softmax((Res(Conv_1(\boldsymbol{F}_{i,s})))^T \times Conv_2(\boldsymbol{F}_{j,s})))) \odot F_{i,s}, \tag{8}$$

where $\odot$ is the element-wise multiplication, $\boldsymbol{F}_{j,s}$ as the output of the self-attention layer represents the feature map of any pedestrian with the same identity as $\boldsymbol{F}_{i,s}$ from different views, and Res represents the reshape operation. *Conv* is a 1×1 convolution. GMP represents the global maximum pooling conducted on all feature channels.

For an image x_j, there is a similar calculation method as follows:

$$\boldsymbol{F}_j = GMP(Res(Softmax((Res(Conv_1(\boldsymbol{F}_{i,s})))^T \times Conv_2(\boldsymbol{F}_{j,s})))^T) \odot F_{i,s}, \tag{9}$$

The above results are summated, and the feature encoder $\boldsymbol{E}$ is optimized by the following cross entropy loss:

$$\boldsymbol{L}_{id} = -\sum_{i=1}^{N} \delta_i \log(\boldsymbol{W}_{id1}(GAP(\boldsymbol{F}_i + \boldsymbol{F}_j))), \tag{10}$$

where $\boldsymbol{W}_{id1}$ is a pedestrian identity classifier corresponding to the integrated features $GAP(\boldsymbol{F}_i + \boldsymbol{F}_j)$. The loss function $\boldsymbol{L}_{id}$ ensures that the pedestrian features extracted after integration are still strongly discriminative. In the above method, x_j only assists x_i in the extraction of non-occluded pedestrian identity features. Under the guidance of self-similarity, the identity-related features in a single pedestrian image are effectively highlighted, and the non-identity-related occlusion and complex background are effectively suppressed. So, the discriminability of pedestrian identity features can be effectively improved under the supervision of person identity labels.

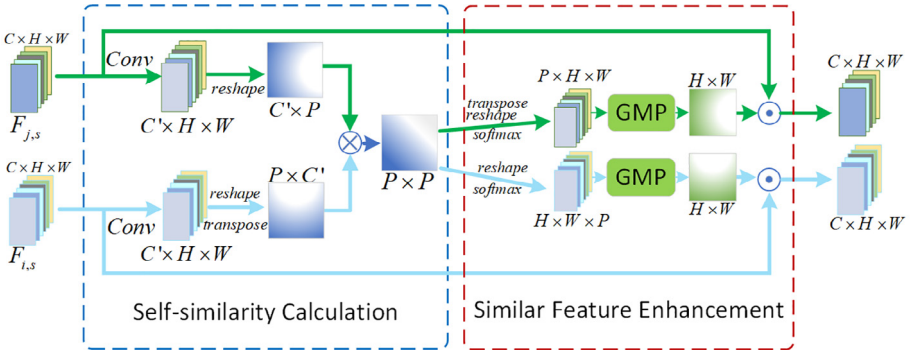

Fig. 7. Self-similarity calculation and similar feature enhancement in SGFDE module. $\otimes$ denotes matrix multiplication. H, W, and C denote the height, width, and number of channels of feature maps, respectively. $C\prime$ represents the number of channels after the convolutional layer (Conv). P is the product of H and W.

3.4. Global Feature Extraction Under Occlusion Suppression

The proposed GFE-UOS module is applied to suppress the impact of occlusions on the enhancement of pedestrian identity features. The fused heat map $\boldsymbol{M}_{f,n} = \boldsymbol{M}_{h,n} + \sum_{k=1}^{K}(\boldsymbol{M}_{l,n}^{k} + \boldsymbol{M}_{r,n}^{k})$ is used to multiply the feature map $\boldsymbol{E}(x_i)$ element by element to filter out the features that are not related to the corresponding areas of pedestrian key points. So, the global feature extraction is achieved as follows:

$$\tilde{\boldsymbol{f}}_i = GAP(\boldsymbol{M}_{f,i} \odot \boldsymbol{E}(x_i)), \tag{11}$$

$$\hat{\boldsymbol{f}}_i = GMP(\boldsymbol{M}_{f,i} \odot \boldsymbol{E}(x_i)), \tag{12}$$

where $\tilde{\boldsymbol{f}}_i$ and $\hat{\boldsymbol{f}}_i$ represent the global features extracted by GAP and the global salient features extracted by GMP, respectively.

To ensure the discriminability of $\tilde{\boldsymbol{f}}_i$ and $\hat{\boldsymbol{f}}_i$, the following loss function is used to optimize the FEN:

$$L_{ice1} = -\sum_{i=1}^{N}\delta_i \log(W_{id2}(\tilde{f}_i)) + \frac{1}{N}\sum_{i=1}^{N}\left[\max_{\tilde{i}} d(\tilde{f}_i, \tilde{f}_{\tilde{i}}) - \min_{i'} d(\tilde{f}_i, \tilde{f}_{i'}) + \tau\right]_+, \tag{13}$$

$$L_{ice2} = -\sum_{i=1}^{N}\delta_i \log(W_{id3}(\hat{f}_i)) + \frac{1}{N}\sum_{i=1}^{N}\left[\max_{\tilde{i}} d(\hat{f}_i, \hat{f}_{\tilde{i}}) - \min_{i'} d(\hat{f}_i, \hat{f}_{i'}) + \tau\right]_+, \tag{14}$$

where $[e]_+ = \max\{e, 0\}$. W_{id2} and W_{id3} are the corresponding person identity classifiers of $\tilde{f}_i$ and $\hat{f}_i$, respectively. $d(\cdot,\cdot)$ represents the Euclidean distance between two feature vectors. $\tilde{f}_{\tilde{i}}$ and $\hat{f}_{\tilde{i}}$ are the corresponding hard positive samples of $\tilde{f}_i$ and $\hat{f}_i$, respectively (hard positive samples are not similar to $\tilde{f}_i$ and $\hat{f}_i$, but have the same pedestrian identities with $\tilde{f}_i$ and $\hat{f}_i$). $\tilde{f}_{i'}$ and $\hat{f}_{i'}$ are the corresponding hard negative samples of $\tilde{f}_i$ and $\hat{f}_i$, respectively (hard negative samples are difficult to distinguish from $\tilde{f}_i$ and $\hat{f}_i$ and have different pedestrian identities from $\tilde{f}_i$ and $\hat{f}_i$). $\tilde{i}$ and i' mean the index of the hard positive sample and the hard negative sample. τ is an interval constant that is set to 0.3 in this paper. In Eqs. (13) and (14), the loss functions L_{ice1} and L_{ice2} ensure that both GAP and GMP can extract discriminative features, respectively. Since $\tilde{f}_i$ and $\hat{f}_i$ have a certain complementarity, $\tilde{f}_i$ and $\hat{f}_i$ are integrated as follows:

$$f_{i,f} = CN(\tilde{f}_i + \hat{f}_i), \tag{15}$$

where $CN(\cdot)$ is short for the cross normalization layer[11]. In Eq. (15), CN is used to prevent the output distribution of the middle layers (including convolution layer, pooling layer, and fully connected layer) of the FEN from changing during the training process. To ensure the strong discriminability of $f_{i,f}$, the cross entropy loss is used to optimize the encoder E as follows:

$$L_{ce} = -\sum_{i=1}^{N}\delta_i \log(W_{id4}(f_{i,f})), \tag{16}$$

where W_{id4} is the pedestrian identity classifier. In the proposed POS, the loss function L_{ce} is used to ensure the discriminability of the features normalized by CN.

3.5. Optimization and Algorithm

The total loss function can be expressed as follows:

$$L = (L_{ice1} + L_{ice2} + L_{ce}) + \lambda_1 L_{cat\_id} + \lambda_2 L_{ped\_id} + \lambda_3 L_{id}, \tag{17}$$

where λ_1, λ_2 and λ_3 are three hyperparameters, which control the corresponding loss functions respectively. ResNet50 pretrained on ImageNet [2] is used as the backbone. Images are adjusted to 256×128 size before inputting to the encoder E. 2,048 feature maps of 16×8 size are finally obtained after inputting to encoder E. All the network parameters are learned by ADM optimizer [9]. The encoder E and pedestrian identity classifiers $W, W_{id1}, W_{id2}, W_{id3}, W_{id4}$, and W_m acting on the m-th key point are trained by minimizing the corresponding loss functions. The optimization process is formalized in Algorithm 1.

Algorithm 1: Key Point-aware Occlusion Suppression and Semantic Alignment for Occluded Person Re-ID

Input: Training images and their labels $D_t = \{x_i, y_i\}_{i=1}^{N_t}$, the number of batches, the maximum number of iterations
Output: The trained E, and the weights w_l, w_r.
for ite1 = 1 **to** *the maximum number of iterations*, **do**:
  for ite2 = 1 **to** *the number of batches*, **do**:
  1: Calculate f_l^k, f_r^k, f_h by Eqs. (1)–(3).
  2: Calculate $\tilde{f}_l^k, \tilde{f}_r^k$ by Eqs. (4) and (5).
  3: Calculate F_i, F_j by Eqs. (8) and (9).
  4: Calculate $\tilde{f}_i, \hat{f}_i$ by Eqs. (11) and (12).
  5: Update E, W, W_m, w_l, w_r by minimizing the loss functions shown in Eqs. (6) and (7).
  6: Update E, W_{id1} by minimizing the loss function shown in Eq. (10).
  7: Update $E, W_{id2}, W_{id3}, W_{id4}$ by minimizing the loss functions shown in Eqs. (13), (14), and (16).
  end for
end for

4. Experiments

4.1. Datasets and Evaluation Metrics

The proposed POS is applied to six challenging datasets to verify its effectiveness and superiority over existing methods. These datasets include the occluded pedestrian re-ID datasets Occluded-DukeMTMC [17] and P-DukeMTMC-reID [49], the partial pedestrian re-ID datasets Partial-REID [42] and Partial-iLIDS [6], and Market-1501 [40] and DukeMTMC-reID [43]. At the same time, the proposed POS is compared with the state-of-the-art methods to verify its superiority, and the effectiveness of each part of POS is verified in ablation experiments. The settings of each dataset are shown in Table 1.

Occluded-DukeMTMC: This dataset is derived from DukeMTMC-reID and designed for occluded person re-ID. The samples in this dataset are collected by eight non-overlapping cameras. The training set contains 15,618 images of 702 pedestrians. The testing set contains 19,871 images of another 519 pedestrians, and each image contains occluded objects.

P-DukeMTMC-reID: The samples in P-DukeMTMC-reID are filtered from DukeMTMC-reID and constructed for occluded person re-ID. All the samples of this dataset are collected by eight non-overlapping cameras. The training set contains 12,927 images of 665 pedestrians. The testing set contains 11,217 images of another 634 pedestrians. In the testing set, 2,163 occluded pedestrian images and the remaining 9,053 non-occluded images compose the Probe and Gallery, respectively.

Partial-REID: This dataset contains 600 images of 60 pedestrians collected by six non-overlapping cameras. Each pedestrian has five full-body images and five partially cropped images of occluded areas. The collections of both partially cropped and full-body images are used as Probe and Gallery, respectively. Due to the small scale of this dataset, Market-1501 and Partial-REID are used as the training and testing sets respectively, which are same as the experimental settings of existing methods [27,4,25,16].

Partial-iLIDS: This dataset is derived from iLiDS [6]. The samples of this dataset come from two non-overlapping cameras, including 238 images of 119 pedestrians in total. 119 images from one camera view are cropped according to occlusions. The other 119 images are full-body images. In this paper, the collections of both cropped and full-body images are used as Probe and Gallery, respectively. In the experiments, Market-1501 and Partial-iLIDS are used as the training and testing sets respectively, which are same as the experimental settings of existing methods [27,4,25,16].

Market-1501: This dataset consists of 32,668 images of 1,501 pedestrians taken by six non-overlapping cameras. In this dataset, the training set contains 12,936 images of 751 pedestrians, and the testing set contains 19,732 images of 750 pedestrians.

Table 1
Settings of different person Re-ID datasets in performance comparison. Ped: Number of pedestrians; Img: Number of images; Cam: Number of cameras.

Datasets	Ped	Training		Gallery (Testing)		Probe (Testing)		Cam
		Ped	Img	Ped	Img	Ped	Img	
Occluded-DukeMTMC	1,221	702	15,618	519	17,661	519	2,210	8
P-DukeMTMC-reID	1,299	665	12,927	634	9,053	634	2,163	8
Partial-REID	60	–	–	60	300	60	300	4
Partial-iLIDS	119	–	–	119	119	119	119	2
Market1501	1,501	751	12,936	750	19,732	750	3,368	6
DukeMTMC-reID	1,404	702	16,522	702	17,661	702	2,228	8

DukeMTMC-reID: This dataset consists of 36,411 images of 1,404 pedestrians taken by eight non-overlapping cameras. In this dataset, the training set contains 16,522 images of 702 pedestrians, and the testing set contains 19,889 images of the remaining 702 pedestrians.

Evaluation Protocol: This paper uses both Cumulative Matching Characteristic (CMC) and mean Average Precision (mAP) [40] as the objective evaluation indicators to evaluate the recognition performance, which are consistent with existing methods [40,11,4].

4.2. Implementation details

In the training process, the same random erasure, random cropping, and horizontal flip as existing solutions [44] are used to achieve data augmentation. In this paper, the batch size N is set to 32. Each batch contains eight pedestrians, and each pedestrian consists of four samples. There are 120 epochs in the whole training process. The initial learning rate of the FEN and all classifiers is set to 0.0002, and the weight decay is set to 0.0005. In the $0 \sim 10$th epoch, the warm-up strategy used in [15] is applied to adjust the learning rate linearly. Starting from the 10th epoch, the learning rate remains unchanged until the 40th epoch. In the 40th epoch, the learning rate attenuates to 10% and remains unchanged until the 70th epoch. In the 70th epoch, the learning rate attenuates to 10% again and remains unchanged until the 120th epoch. In the experiments, the hyperparameters λ_1, λ_2 and λ_3 are set to 0.5, 0.7, and 0.4, respectively. SubSection 4.5 shows the analysis process of these three hyperparameters. The experiments in this paper are implemented on a single NVIDIA GeForce RTX2080Ti GPU, and the pytorch framework [18] is used. In the testing process, the cosine distance is used to measure the similarity.

4.3. Ablation Study

The proposed POS is primarily composed of three modules, i.e., KPA-SA, SGFDE, and GFE-UOS. In the proposed POS, the Baseline is obtained by minimizing loss L_{ice1} through training. To verify the effectiveness of each module, three different modules are added to the Baseline in turn, the corresponding combinations of different modules and Baseline are compared with each other. During this process, all the experiments are performed on Occluded-DukeMTMC, and the corresponding results are shown in Table 2.

The effectiveness of KPA-SA: In this paper, the KPA-SA module is mainly used to alleviate the weak discriminability of features caused by the spatial position misalignment of pedestrian images. As shown in Table 2, after adding KPA-SA to the Baseline, Baseline + KPA-SA increases the corresponding recognition rates/accuracy 4.9%, 5.7%, 6.1%, and 2.8% on Rank-1, Rank-5, Rank-10, and mAP, respectively. The experimental results confirm KPA-SA can improve the model performance. KPA-SA achieves the semantic alignment of pedestrian features; thus, the interference caused by the posture changes and occlusion is reduced.

The effectiveness of SGFDE: The SGFDE is proposed to further improve the discriminability of features. As shown in Table 2, Baseline + SGFDE also increases the corresponding recognition rates/accuracy of Baseline 3.6%, 3.9%, 3.8%, and 3.6% on Rank-1, Rank-5, Rank-10, and mAP, respectively. The experimental results confirm SGFDE can improve the model performance.

The effectiveness of GFE-UOS: The GFE-UOS uses both GMP and GAP to avoid the loss of pedestrian saliency information, and also considers the interference caused by the occlusion on feature extraction. As shown in Table 2, Baseline + GFE-UOS also increases the corresponding recognition rates/accuracy of Baseline 8.3%, 10.4%, 10.6%, and 1.0% on Rank-1, Rank-5, Rank-10, and mAP, respectively. The experimental results confirm that GFE-UOS can play a positive role in improving the model performance.

The effectiveness of two modules: As shown in Table 2, the overall performance of Baseline + two modules is better than the corresponding performance of Baseline + a single module. Specifically, Baseline + KPA-SA + SGFDE and Baseline + SGFDE + GFE-UOS obtain better recognition rates on Rank-1, Rank-5, Rank-10, and mAP than all Baseline + a single module. Baseline + KPA-SA + SGFDE is optimized by minimizing $L_{ice1} + \lambda_1 L_{cat\_id} + \lambda_2 L_{ped\_id} + \lambda_3 L_{id}$. After adding SGFDE to Baseline + KPA-SA, the corresponding recognition rates/accuracy increase 1.5%, 0.9%, and 2.6% on Rank-1, Rank-10, and

Table 2
Ablation study. The corresponding CMC rates(%) are listed to show the impact of each module on the proposed method.

Methods	Rank-1	Rank-5	Rank-10	mAP
Baseline	49.0	64.3	69.8	42.6
Baseline + KPA-SA	53.9	70.0	75.9	45.4
Baseline + SGFDE	52.6	68.2	73.6	46.2
Baseline + GFE-UOS	57.3	74.7	80.4	43.6
Baseline + KPA-SA + SGFDE	55.4	70.0	76.8	48.0
Baseline + KPA-SA + GFE-UOS	63.5	78.1	83.5	53.6
Baseline + SGFDE + GFE-UOS	61.6	78.1	83.7	53.2
Baseline + KPA-SA + SGFDE + GFE-UOS	65.0	79.0	83.8	54.0

mAP, respectively. Similarly, the SGFDE can also improve the performance of Baseline + GFE-UOS. The comparative results further confirm that SGFDE plays a positive role in enhancing the discriminability of features.

The effectiveness of three modules: As a complete model, Baseline + KPA-SA + SGFDE + GFE-UOS is trained by minimizing the loss $(L_{ice1} + L_{ice2} + L_{ce}) + \lambda_1 L_{cat\_id} + \lambda_2 L_{ped\_id} + \lambda_3 L_{id}$. Compared with all the Baseline, Baseline + a single module, and Baseline + two modules, Baseline + KPA-SA + SGFDE + GFE-UOS achieves the best performance on Rank-1, Rank-5, Rank-10, and mAP as shown in Table 2. So, the efficiency of the propose method is confirmed.

Performance analysis of each module on full-image datasets: Table 3 shows the performance of each module of the proposed method on full-image datasets. When the proposed method is applied to two full-image datasets, DukeMTMC-reID and Market-1501, each module still plays a positive role in improving recognition performance. Compared with the corresponding performance on occlusion datasets, the performance improvement of each module is quite limited. As the main reason, the SGFDE and KPA-SA modules are designed for occlusion images. Image information is almost complete in full images, so the SGFDE and KPA-SA modules are difficult to achieve the expected performance.

The ablation study on the number of key points: Fig. 8 shows the effect of the different number of key points on heat maps and the performance of the proposed POS. As shown in Fig. 8 (a), when only one key point of the head is used, POS can focus on the area corresponding to the key point. When 13 key points covering the whole pedestrian body are used, POS can focus on all non-occluded areas of the pedestrian body. The coverage of key points on the pedestrian body determines the performance of POS on the extraction of appearance features. As shown in Fig. 8 (b), when the number of key points increases, the recognition accuracy on Rank-1 is gradually improved until the number of key points reaches 13. Additionally, the recognition accuracy on mAP is also slightly improved. The above experimental results confirm complete pedestrian appearance features are conducive to pedestrian identity matching. When there are too many key points, the semantic information of each key point cannot be accurately distinguished due to the overlap between the semantics of each key point, which affects the semantic alignment of global feature channels. Therefore, when the number of key points $M = 17$, the recognition performance of the proposed model cannot be further improved, and even mAP has a downward trend.

4.4. Comparison with The State-of-the-art Methods

Results on Occluded-DukeMTMC: To verify the performance of the proposed POS, it is applied to the commonly used occlusion dataset Occluded-Duke, and its performance is compared with the state-of-the-art methods, including PartAlign [38], PCB [24], PartBili [21], DSR [6], AdOccl [8], PGFA [17], HOReID [27], SORN [36] and MHSANet [25]. As shown in Table 4, since the first seven methods do not consider the impact of occlusion on recognition performance, the recognition rate on Rank-1 and accuracy on mAP are low. In this type of method, both 42.6% recognition rate on Rank-1 and 33.7% accuracy on mAP obtained by PCB are the best results, which mean occlusion is a key factor affecting recognition performance.

AdOccl, PGFA, HOReID, and MHSANet consider the impact of occlusions on recognition performance. Therefore, compared with the first type of method, the recognition rates on Rank-1 and accuracy on mAP of the four methods are significantly improved. Both the latest methods HOReID and MHSANet achieve more than 55% recognition rate on Rank-1 and 42% accuracy on mAP. Compared with the above methods, the proposed POS considers the semantic alignment of pedestrian features, uses the symmetry of the pedestrian's body to compensate for the information loss of the occluded areas, and utilizes the self-similarity of pedestrian identity features to enhance the discriminability of features, so it can achieve a better matching accuracy. Specifically, the proposed POS achieves 65.0% recognition rate on Rank-1 and 54.0% accuracy on mAP, respectively. Compared with MHSANet (SONR), the recognition accuracy and accuracy of the proposed POS on Rank-1 and mAP increase by 9.6% (7.4%) and 11.6% (7.7%), respectively. All the above results confirm that the effectiveness of the proposed POS and its superiority over existing methods.

Results on P-DukeMTMC-reID: To further verify the performance under occlusion interference, the proposed model is applied to P-DukeMTMC-reID, and its performance is compared with the latest nine methods, including HACNN [13], PartBili [21], PCB [24], OSNet [45], PGFA [17], TCSDO [50], PVPM [4], and MHSANet [25]. As shown in Table 5, the first five methods do not consider the interference caused by occlusions, so the corresponding recognition rates are relatively low. TCSDO and

Table 3
Performance analysis of each module on DukeMTMC-reID and Market-1501. The corresponding CMC rates(%) are listed to show the impact of each module on the full-image datasets.

Methods	Market-1501			DukeMTMC-reID		
	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP
Baseline	93.6	98.1	85.2	86.3	94.3	75.1
Baseline + KPA-SA	94.0	98.0	85.9	87.9	94.3	76.1
Baseline + SGFDE	94.2	98.1	85.7	87.9	94.3	76.6
Baseline + GFE-UOS	94.8	98.3	86.0	88.1	94.6	76.2
Baseline + KPA-SA + SGFDE	94.8	98.2	86.1	88.0	94.3	76.7
Baseline + KPA-SA + GFE-UOS	94.9	98.2	85.9	88.6	94.6	76.7
Baseline + SGFDE + GFE-UOS	94.9	98.2	86.1	88.2	94.6	76.6
Baseline + KPA-SA + SGFDE + GFE-UOS	95.0	98.3	86.2	88.7	94.6	76.7

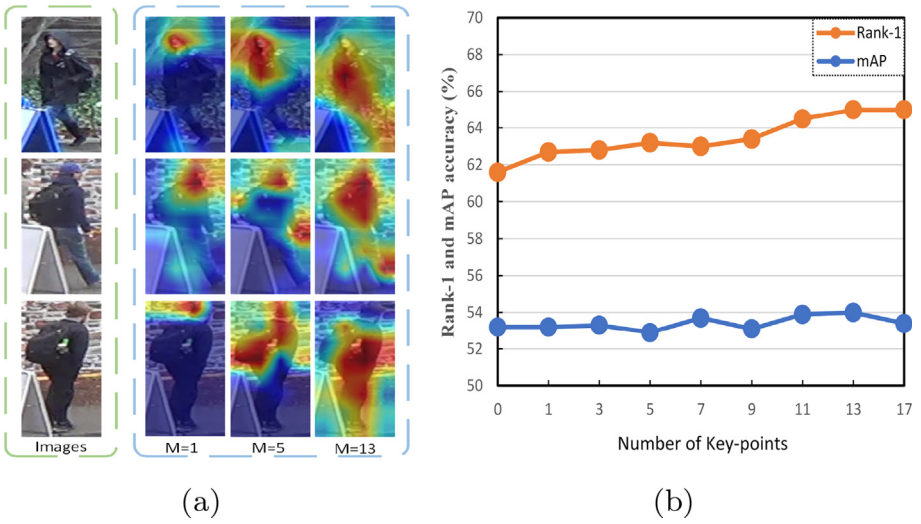

Fig. 8. Ablation study on the number of key points. (a) Effect of the different number of key points on heat maps, (b) Effect of the different number of key points on Rank-1 and mAP accuracy.

Table 4
Comparison of the proposed POS with the state-of-the-art methods on Occluded-DukeMTMC. The CMC rates(%) are listed. '–' denotes that no reported result is available. The best results are shown in bold and the second-best results are shown in italic.

Methods	Rank-1	Rank-5	Rank-10	mAP
PartAlign[38]	28.8	44.6	51.0	20.2
PCB[24]	42.6	57.1	62.9	33.7
PartBili[21]	36.9	–	–	–
DSR[6]	40.8	58.2	65.2	30.4
AdOccl[8]	44.5	–	–	32.2
PGFA[17]	51.4	68.6	74.9	37.3
HOReID[27]	55.1	–	–	43.8
SORN[36]	*57.6*	*73.3*	*79.0*	*46.3*
MHSANet[25]	55.4	70.2	76.4	42.4
POS(Proposed)	**65.0**	**79.0**	**83.8**	**54.0**

Table 5
Comparison of the proposed POS with the state-of-the-art methods on P-DukeMTMC-reID. The CMC rates(%) are listed. '–' denotes that no reported result is available. The best results are shown in bold and the second-best results are shown in italic.

Methods	Rank-1	Rank-5	Rank-10	mAP
HACNN[13]	30.4	42.1	49.0	17.0
PartBili[21]	39.2	50.6	56.4	25.4
PCB[24]	40.4	54.6	61.1	23.4
OSNet[45]	33.7	46.5	54.0	20.1
PGFA[17]	44.2	56.7	63.0	23.1
TCSDO[50]	51.4	58.5	69.7	*55.6*
PVPM[4]	51.5	64.4	69.6	29.2
MHSANet[25]	*67.9*	*79.7*	*83.7*	37.6
POS(Proposed)	**72.1**	**84.2**	**87.7**	**59.4**

PGFA consider the interference caused by occlusions to pedestrian identity matching, so the corresponding recognition rates on Rank-1 are improved compared with the first five methods. Similarly, the accuracy on mAP obtained by TCSDO is also improved and higher than the corresponding ones obtained by the first five methods. The accuracy on mAP obtained by PGFA is higher than the corresponding ones obtained by HACNN, and OSNet, but only lower than the corresponding ones obtained by PartBili and PCB. The latest occluded person re-ID methods TCSDO, PVPM, and MHSANet achieve more than 51% recognition rate on Rank-1 and 29.2% accuracy on mAP, respectively. The proposed POS achieves 72.1% recognition rate on Rank-1 and 59.4% accuracy on mAP, respectively. Compared with MHSANet (PVPM), the proposed POS increases the recognition rate on Rank-1 and accuracy on mAP by 4.2% (20.6%) and 21.8% (30.2%), respectively. The above results further confirm the effectiveness of the proposed POS in occluded person re-ID and its superiority over existing methods.

Results on Partial-REID and Partial-iLIDS: The proposed POS is applied to two partial-ReID datasets, Partial-REID and Partial-iLIDS, to further verify its performance by comparing with AMCSWM[42], IDE [40], DSR [6], PCB [24], OPR [49], OSNet [45], PGFA [17], FPR [7], TCSDO [50], HOReID [27], PVPM [4], MHSANet [25] and STNReID [16]. Table 6 shows the recognition accuracy of different methods on Rank-1, Rank-3, and mAP. For the Partial-REID (Partial-iLIDs) dataset, the proposed POS obtains 86.1% (76.1%) recognition rate on Rank-1, 91.3% (88.5%) recognition rate on Rank-3, and 80.9% (70.3%) accuracy on mAP, respectively. For the Partial-REID dataset, compared with the second-best result obtained by HOReID, this proposed method improves the recognition accuracy on Rank-1 by 0.8%. For the Partial-iLIDs dataset, the proposed POS improves the recognition accuracy on Rank-1 by 2.5 % compared with the second-best accuracy obtained by HOReID. It means that the proposed POS can also achieve better recognition performance on partial datasets.

Results on non-occluded datasets: Although the three modules: KPA-SA, SGFDE and GFE-UOS in the proposed method are all designed for occluded person re-ID, these three modules do not have any side effect on the corresponding recognition performance of the proposed method, when a/multiple target pedestrians are not occluded. Therefore, the proposed method can be applied to non-occluded person re-ID. In the following experiments, two challenging non-occluded datasets Market-1501 and DukeMTMC-reID are used to verify the performance of the proposed POS by comparing with MCAM [20], PCBRPP [24], FANN [47], PGFA [17], AANet [26], CAMA [32], OSNet [45], HACNDHA [31], SORN [36], HOReID [27] and MHSANet [25]. Table 7 shows the performance of different methods on Market-1501 and DukeMTMC-reID. For Market-1501 and DukeMTMC-reID, the proposed POS achieves the 95.0% and 88.7% recognition rates on Rank-1, and 86.2% and 76.7% accuracy on mAP, respectively. Compared with the second-best result obtained by OSNet and the performance of the latest methods HOReID and MHSANet, the proposed POS shows stronger competitiveness. It further confirms that the effectiveness of the proposed POS is superior to the existing similar methods.

4.5. Parameter Analysis

The proposed POS contains hyper-parameters $\lambda_1, \lambda_2, \lambda_3$. The impact of the hyper-parameters λ_1, λ_2 and λ_3 on model performance is analyzed on Occluded-DukeMTMC. One of three hyper-parameters is changed in the analysis process, while keeping two remaining hyper-parameters unchanged.

The impact of parameter λ_1 on model performance: Fig. 9(a) shows the impact of λ_1 on model performance, when the value of λ_1 increases from 0.1 to 2. When $\lambda_1 \in [0.3, 0.5]$, the proposed POS shows better performance. When $\lambda_1 = 0.5$, the proposed POS achieves the highest recognition rate on Rank-1 and accuracy on mAP, respectively. When $\lambda_1 > 0.5$, the model performance decreases. Therefore, it is reasonable to set $\lambda_1 = 0.5$ in this paper.

The impact of parameter λ_2 on model performance: In Eq. (17), λ_2 is used to adjust L_{ped_id}. Fig. 9(b) shows the impact of different values of λ_2 on model performance. When λ_2 increases from 0 to 0.2, the recognition accuracy of the model on Rank-1 gradually decreases, and reaches the lowest value at $\lambda_2 = 0.2$. When the value of λ_2 increases from 0.2 to 0.7, the model achieves the highest recognition rate on Rank-1 and accuracy on mAP. So, $\lambda_2 = 0.7$ is a reasonable setting.

The impact of parameter λ_3 on model performance: In Eq. (11), λ_3 is used to control L_{id}. Fig. 9(c) shows the model performance according to the value changes of λ_3. When $\lambda_3 = 0.4$, the proposed POS achieves the highest recognition performance on Rank-1 and accuracy on mAP, respectitvely. When $\lambda_3 > 0.4$, the performance of the proposed POS decreases. Therefore, λ_3 is set to 0.4 in the experiments.

Table 6
Comparison of the proposed POS with the state-of-the-art methods on Partial-REID and Partial-iLIDS. The CMC rates(%) are listed. "–" denotes that no reported result is available. The best results are shown in bold and the second-best results are shown in italic.

Methods	Partial-REID			Partial-iLIDS		
	Rank-1	Rank-3	mAP	Rank-1	Rank-3	mAP
AMCSWM[42]	37.3	46.0	31.3	21.0	32.8	–
IDE[40]	51.7	–	52.4	–	–	–
DSR[6]	50.7	70.0	–	58.8	67.2	–
PCB[24]	56.3	–	54.7	46.8	–	40.2
OPR[49]	78.5	–	–	–	–	–
OSNet[45]	48.7	–	49.3	–	–	–
PGFA[17]	68.0	80.0	–	69.1	80.9	–
FPR[7]	81.0	–	*76.6*	68.1	–	*61.8*
TCSDO[50]	69.2	–	73.1	–	–	–
HOReID[27]	*85.3*	*91.0*	–	72.6	*86.4*	–
PVPM[4]	78.3	–	–	–	–	–
MHSANet[25]	81.3	87.7	–	*73.6*	85.4	–
STNReID[16]	66.7	80.3	–	54.6	71.3	–
POS(Proposed)	**86.1**	**91.3**	**80.9**	**76.1**	**88.5**	**70.3**

Table 7
Comparison of the proposed POS with the state-of-the-art methods on non-occluded Market-1501 and DukeMTMC-reID. The CMC rates(%) are listed. "–" denotes that no reported result is available. The best results are shown in bold and the second-best results are shown in italic.

Methods	Market-1501		DukeMTMC-reID	
	Rank-1	mAP	Rank-1	mAP
MCAM[20]	83.8	74.3	–	–
PCBRPP[24]	93.8	81.6	83.3	69.2
FANN[47]	90.3	76.1	–	–
PGFA[17]	91.2	76.8	82.6	65.5
AANet[26]	93.9	83.4	87.7	*74.3*
CAMA[32]	94.7	84.5	85.8	72.9
OSNet[45]	*94.8*	*84.9*	*88.6*	73.5
HACNDHA[31]	91.3	76.0	81.3	64.1
SORN[36]	94.8	84.5	86.9	74.1
HOReID[27]	94.2	84.9	86.9	75.6
MHSANet[25]	94.6	84.0	87.3	73.1
POS(Proposed)	**95.0**	**86.2**	**88.7**	**76.7**

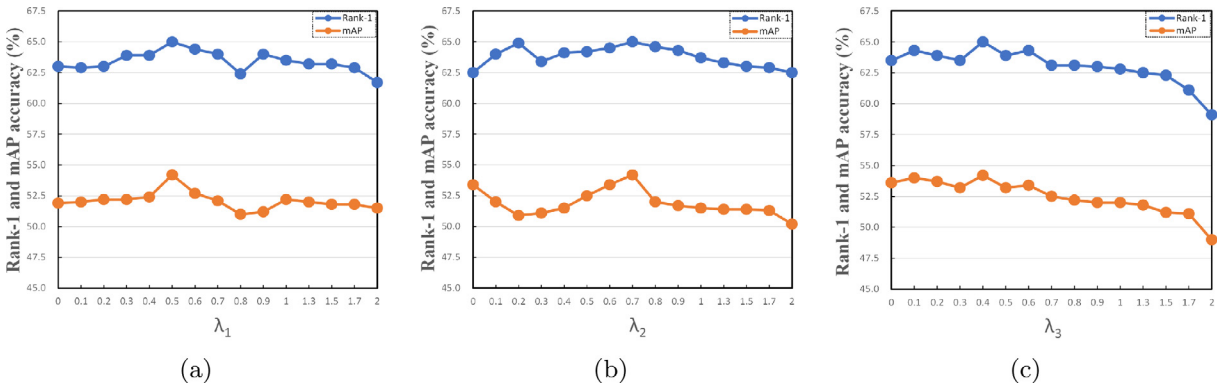

Fig. 9. Effect analysis on hyper-parameters λ_1, λ_2 and λ_3 of the proposed POS. When one parameter is evaluated, other parameters are fixed at the optimal values. (a) The effect of λ_1 on Rank-1 and mAP accuracy, (b) The effect of λ_2 on Rank-1 and mAP accuracy, (c) The effect of λ_3 on Rank-1 and mAP accuracy.

4.6. Robustness Analysis to Occlusion

Both the proposed method and HOReID are developed based on key point detection. Different degrees of occlusion bring different levels of difficulty to the detection of pedestrian key points. To verify the performance of the proposed method under different degrees of occlusion, 25%, 33.3%, 50%, 66.7%, and 75% occlusion are applied to the query samples of DukeMTMC-reID dataset. The main obstructions include umbrellas, billboards, vehicles, and so on. Fig. 10 shows the performance of different models on DukeMTMC-reID dataset with different degrees of occlusions. Fig. 11 visualizes the impact of different degrees of occlusions on the retrieval results of the proposed method.

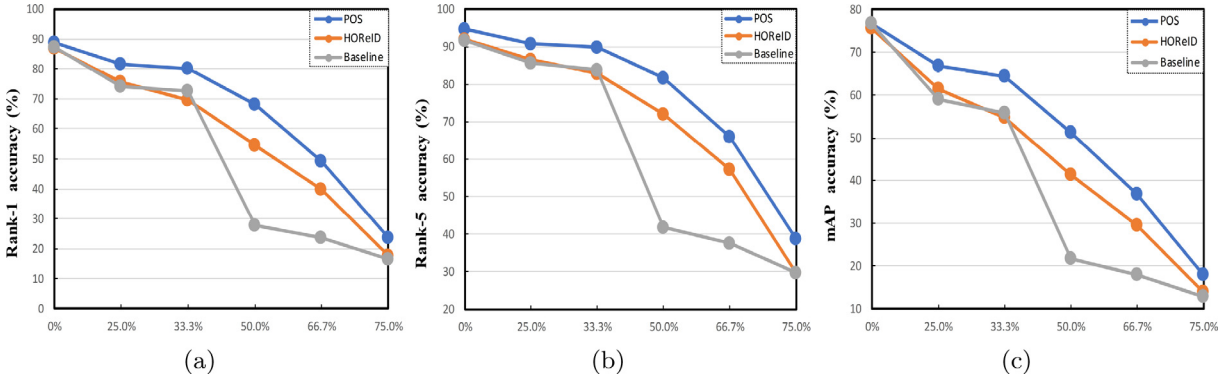

Fig. 10. Influence of different levels of occlusions on pedestrian identity matching.

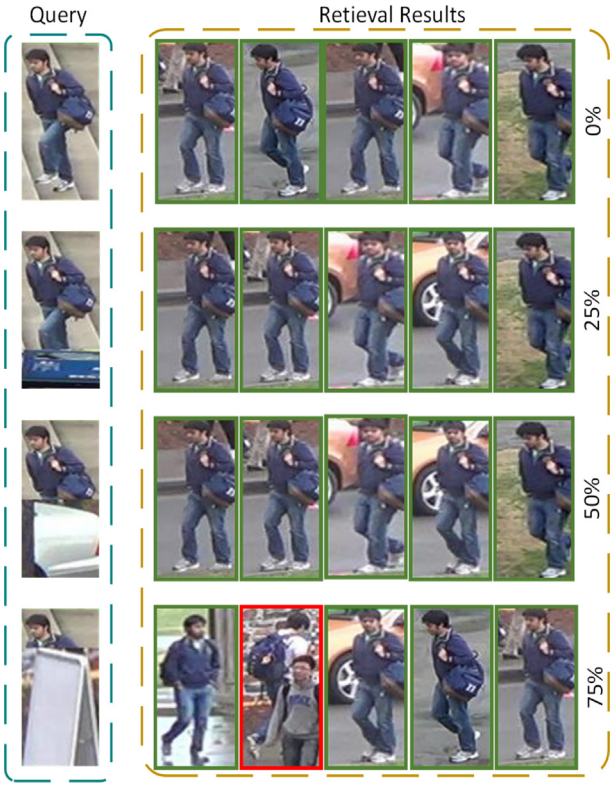

Fig. 11. The top five retrieval results of the proposed method under different degrees of occlusion. The green and red frames indicate correct and incorrect matching results, respectively.

According to the results shown in Fig. 10, as the degree of occlusions gradually increases, the difficulty of key point detection increases and the performance of the proposed method decreases. Compared with baseline and HOReID specially designed for the re-ID of occluded pedestrians, the proposed POS achieves better performance. So, it confirms that the proposed model is more robust to occlusion than the other two comparative methods. Fig. 11 shows the retrieval results of POS on occlusion images with different degrees. The proposed method has relatively stable performance even under a high degree of occlusion.

4.7. Training Cost Analysis

Table 8 shows the training cost, test cost, parameters and FLOPs of Baseline, HOReID, DSR and POS on Occluded-DukeMTMC and Market-1501. Compared with Baseline, POS improves the recognition accuracy of Rank-1 and mAP by 16% and 11.4% respectively with an increase of 19.8G FLOPs, 63.6 M parameters and 5.9 h running time on Occluded-DukeMTMC. Compared with Baseline, the training time, Rank-1, and mAP of POS increase by 5.2 h, 0.5%, and 0.3% respectively on Market-1501. Compared with HOReID based on key point detection, POS spent 1.8 h and 1.2 h more training time respectively on Occluded-DukeMTMC and Market-1501. However, POS improved Rank-1 and mAP by 9.9% and 10.2% (0.8% and 1.3%) respectively on Occluded-DukeMTMC (Market-1501). Moreover, the testing time consumed by POS on both Occluded-DukeMTMC and Market-1501 is less than that of HOReID. Compared with DSR without key point detection,

Table 8
Training time and testing time analysis of different methods on Occluded-DukeMTMC and Market-1501. "–" indicates that no reported result is available.

Methods	Params/M	FLOPs/G	Occluded-DukeMTMC				Market-1501			
			Training Time/h	Testing Time/s	Rank-1	mAP	Training Time/h	Testing Time/s	Rank-1	mAP
Baseline	23.0	4.0	1.6	77	49.0	42.6	1.5	115	94.5	85.9
DSR	23.1	6.2	2.5	89	52.1	44.5	2.1	130	83.6	64.3
HOReID	–	–	5.7	206	55.1	43.8	5.5	252	94.2	84.9
POS	86.6	23.8	7.5	198	65.0	54.0	6.7	241	95.0	86.2

Table 9
Effect of CCC.

Datasets	POS		POS + CCC	
	Rank-1	mAP	Rank-1	mAP
Occluded-DukeMTMC	65.0	54.0	65.0	54.2
DukeMTMC-reID	88.7	76.7	88.9	77.1
Market-1501	95.0	86.2	95.0	87.0

POS consumed more training and testing time, but it achieved better performance. Since the proposed POS can be pre-trained before actual deployment, there is no need to train it during practical use. Therefore, the increase in model training cost does not affect the practical deployment and application of the proposed POS. Owing to the introduction of key point detection, the testing efficiency of the propose POS decreases, but it can still meet the needs of real-world applications.

4.8. Discussion

Inspired by [34], the cross fusion mechanism discussed in Section 3.2 is improved by introducing cycle consistency constraints (CCC). Specifically, let FC_{lr} and FC_{rl} be two full connection layers, $f_r^{k'} = FC_{lr}(f_l^k), f_l^{k'} = FC_{rl}(f_r^k), f_r^{k''} = FC_{lr}(f_l^{k'})$, and $f_l^{k''} = FC_{rl}(f_r^{k'})$. So, the corresponding loss of CCC is $\sum_{k=1} \|f_r^{k''} - f_r^k\|_2^2 + \|f_l^{k''} - f_l^k\|_2^2$, which can ensure the consistency of identity and semantic information in the transformation process and reduce information loss. The experimental results shown in Table 9 demonstrate the effect of CCC.

5. Conclusion

This paper proposes an occluded person re-ID method called POS consisting of KPA-SA, SGFDE, and GFE-UOS three modules. The proposed KPA-SA can solve the misalignment of pedestrian images in spatial positions and use the paired key points of the human body to alleviate the impact of occlusions. The proposed SGFDE utilizes the similarity of the same pedestrian captured from different views to realize the occlusion suppression by the FEN and enhancement of pedestrian identity features. The proposed GFE-UOS uses the fused heat maps of different key points to extract the features of non-occluded areas, which enhances the discriminability of global features. This paper verifies the effectiveness of each module of the proposed POS in the ablation experiments. In addition, the comparative experimental results on six public datasets verify the effectiveness of the proposed POS and its superiority over the state-of-the-art methods. In future, person re-ID across multiple occlusion datasets will be further explored to promote the popularization and application of person re-ID in real-world scenes by integrating the idea of feature extraction in [1].

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 61966021, Grant 61772455, Grant 61562053, Grant 61962032, National Key Research and Development Plan Project under Grant 2018YFC0830105 and Grant 2018YFC0830100.

References

[1] Xiaojun Chang, Feiping Nie, Sen Wang, Yi Yang, Xiaofang Zhou, Chengqi Zhang, Compound rank-k projections for bilinear analysis, IEEE Transactions on Neural Networks and Learning Systems 27 (7) (2016) 1502–1513.
[2] Jia Den, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, Fei-Fei Li, Imagenet: A large-scale hierarchical image database, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009, pp. 248–255.
[3] Changxing Ding, Kan Wang, Pengfei Wang, Dacheng Tao, Multi-task learning with coarse priors for robust part-aware person re-identification, IEEE Transactions on Pattern Analysis and Machine Intelligence 44 (3) (2022) 1474–1488.
[4] Shang Gao, Jingya Wang, Lu. Huchuan, Zimo Liu, Pose-guided visible part matching for occluded person re-id, in: The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11744–11752.
[5] Yi Hao, Nannan Wang, Xinbo Gao, Jie Li, Xiaoyu Wang, Dual-alignment feature embedding for cross-modality person re-identification, in: Proceedings of the ACM International Conference on Multimedia (ACMMM), 2019, pp. 57–65.
[6] Lingxiao He, Jian Liang, Haiqing Li, Zhenan Sun, Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7073–7082.
[7] Lingxiao He, Yinggang Wang, Wu Liu, He Zhao, Zhenan Sun, and Jiashi Feng. Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification. In the IEEE International Conference on Computer Vision (ICCV), pages 8450–8459, 2019..

[8] Houjing Huang, Dangwei Li, Zhang Zhang, Xiaotang Chen, Kaiqi Huang, Adversarially occluded samples for person re-identification, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5098–5107.
[9] Diederik Kingma, Jimmy Ba, Adam: A method for stochastic optimization, in: International Conference for Learning Representations (ICLR), 2015.
[10] Huafeng Li, Yiwen Chen, Dapeng Tao, Yu. Zhengtao, Guanqiu Qi, Attribute-aligned domain-invariant feature learning for unsupervised domain adaptation person re-identification, IEEE Transactions on Information Forensics and Security 16 (2021) 1480–1494.
[11] Huafeng Li, Neng Dong, Zhengtao Yu, Dapeng Tao, Guanqiu Qi, Triple adversarial learning and multi-view imaginative reasoning for unsupervised domain adaptation person re-identification, in: IEEE Transactions on Circuits and Systems for Video Technology, 2021, 1–1.
[12] Huafeng Li, Jian Pang, Dapeng Tao, Yu. Zhengtao, Cross adversarial consistency self-prediction learning for unsupervised domain adaptation person re-identification, Information Sciences 559 (2021) 46–60.
[13] Wei Li, Xiatian Zhu, Shaogang Gong, Harmonious attention network for person re-identification, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2285–2294.
[14] Yuanyuan Li, Sixin Chen, Guanqiu Qi, Zhiqin Zhu, Matthew Haner, Ruihua Cai, A gan-based self-training framework for unsupervised domain adaptive person re-identification, Journal of Imaging 7 (4) (2021) 62.
[15] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2019..
[16] Hao Luo, Wei Jiang, Xing Fan, Chi Zhang, Stnreid: Deep convolutional networks with pairwise spatial transformer networks for partial person re-identification, IEEE Transactions on Multimedia 22 (11) (2020) 2905–2913.
[17] Yu.Wu. Jiaxu Miao, Ping Liu, Yuhang Ding, Yi Yang, Pose-guided feature alignment for occluded person re-identification, in: The IEEE International Conference on Computer Vision (ICCV), 2019, pp. 542–551.
[18] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. volume 32, pages 8026–8037, 2019..
[19] Guanqiu Qi, Hu. Gang, Xiaofei Wang, Neal Mazur, Zhiqin Zhu, Matthew Haner, Exam: A framework of learning extreme and moderate embeddings for person re-id, Journal of Imaging 7 (1) (2021) 6.
[20] Chunfeng Song, Yan Huang, Wanli Ouyang, Liang Wang, Mask-guided contrastive attention model for person re-identification, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1179–1188.
[21] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. In the European Conference on Computer Vision (ECCV), pages 402–419, 2018..
[22] K. Sun, B. Xiao, D. Liu, J. Wang, Deep high-resolution representation learning for human pose estimation, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5686–5696.
[23] Yifan Sun, Qin Xu, Yali Li, Chi Zhang, Yikang Li, Shengjin Wang, and Jian Sun. Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. In the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 393–402, 2019..
[24] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, Shengjin Wang, Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline), in: The European Conference on Computer Vision (ECCV), 2018, pp. 480–496.
[25] Hongchen Tan, Xiuping Liu, Shengjing Tian, Baocai Yin, and Xin Li. Mhsa-net: Multi-head self-attention network for occluded person re-identification. ArXiv:2008.04015, 2020..
[26] Chiat-Pin Tay, Sharmili Roy, Kim-Hui Yap, Aanet: Attribute attention network for person re-identifications, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7134–7143.
[27] Guan'an Wang, Shuo Yang, Huanyu Liu, Zhicheng Wang, Yang Yang, Shuliang Wang, Gang Yu, Erjin Zhou, and Jian Sun. High-order information matters: Learning relation and topology for occluded person re-identification. In the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 6449–6458, 2020..
[28] Jianchen Wang, Liming Yuan, Xu. Haixia, Gengsheng Xie, Xianbin Wen, Channel-exchanged feature representations for person re-identification, Information Sciences 562 (2021) 370–384.
[29] Kai Wang, Shichao Dong, Nian Liu, Junhui Yang, Tao Li, Hu. Qinghua, Pa-net: Learning local features using by pose attention for short-term person re-identification, Information Sciences 565 (2021) 196–209.
[30] Xiaolong Wang, Ross Girshick, Abhinav Gupta, Kaiming He, Non-local neural networks, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7794–7803.
[31] Zheng Wang, Junjun Jiang, Wu. Yang, Mang Ye, Xiang Bai, Shin'ichi Satoh, Learning sparse and identity-preserved hidden attributes for person re-identification, IEEE Transactions on Image Processing 29 (1) (2019) 2013–2025.
[32] Wenjie Yang, Houjing Huang, Zhang Zhang, Xiaotang Chen, Kaiqi Huang, and Shu Zhang. Towards rich feature discovery with class activation maps augmentation for person re-identification. In the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1389–1398, 2019..
[33] Xi Yang, Liangchen Liu, Nannan Wang, Xinbo Gao, A two-stream dynamic pyramid representation model for video-based person re-identification, IEEE Transactions on Image Processing 30 (2021) 6266–6276.
[34] Di Yuan, Xiaojun Chang, Po-Yao Huang, Qiao Liu, Zhenyu He, Self-supervised deep correlation tracking, IEEE Transactions on Image Processing 30 (2021) 976–985.
[35] Guoqing Zhang, Junchuan Yang, Yuhui Zheng, Ye Wang, Wu. Yi, Shengyong Chen, Hybrid-attention guided network with multiple resolution features for person re-identification, Information Sciences 578 (2021) 525–538.
[36] Xiaokang Zhang, Yan Yan, Jing-Hao Xue, Yang Hua, and Hanzi Wang. Semantic-aware occlusion-robust network for occluded person re-identification. IEEE Transactions on Circuits and Systems for Video Technology, pages 2764–2778, 2021..
[37] Yulin Zhang, Bo Ma, Yuqing Feng, Meng Li, Pmt-net: Progressive multi-task network for one-shot person re-identification, Information Sciences 568 (2021) 133–146.
[38] Liming Zhao, Xi Li, Yueting Zhuang, Jingdong Wang, Deeply-learned part-aligned representations for person re-identification, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3219–3228.
[39] Shizhen Zhao, Changxin Gao, Jun Zhang, Hao Cheng, Chuchu Han, Xinyang Jiang, Xiaowei Guo, Wei-Shi Zheng, Nong Sang, and Xing Sun. Do not disturb me: Person re-identification under the interference of other pedestrians. In the European Conference on Computer Vision (ECCV), pages 647–663, 2020..
[40] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In the IEEE International Conference on Computer Vision (ICCV), pages 1116–1124, 2015..
[41] Wei-Shi Zheng, Shaogang Gong, Tao Xiang, Person re-identification by probabilistic relative distance comparison, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 649–656.
[42] Wei-Shi Zheng, Xiang Li, Tao Xiang, Shengcai Liao, Jianhuang Lai, and Shaogang Gong. Partial person re-identification. In the IEEE International Conference on Computer Vision (ICCV), pages 4678–4686, 2015..
[43] Zhedong Zheng, Liang Zheng, Yi Yang, Unlabeled samples generated by gan improve the person re-identification baseline in vitro, in: The IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3774–3782.
[44] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, Yi Yang, Random erasing data augmentation, in: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2020, pp. 13001–13008.
[45] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, Tao Xiang, Omni-scale feature learning for person re-identification, in: The IEEE International Conference on Computer Vision (ICCV), 2019, pp. 3702–3712.
[46] Runwu Zhou, Xiaojun Chang, Lei Shi, Yi-Dong Shen, Yi Yang, Feiping Nie, Person reidentification via multi-feature fusion with adaptive graph learning, IEEE Transactions on Neural Networks and Learning Systems 31 (5) (2020) 1592–1601.

[47] Sanping Zhou, Jinjun Wang, Deyu Meng, Yudong Liang, Yihong Gong, Nanning Zheng, Discriminative feature learning with foreground attention for person re-identification, IEEE Transactions on Image Processing 28 (9) (2019) 4671–4684.
[48] Zhiqin Zhu, Yaqin Luo, Sixin Chen, Guanqiu Qi, Neal Mazur, Chengyan Zhong, Qiwang Li, Camera style transformation with preserved self-similarity and domain-dissimilarity in unsupervised person re-identification, Journal of Visual Communication and Image Representation 80 (2021) 103303.
[49] Jiaxuan Zhuo, Zeyu Chen, Jianhuang Lai, Guangcong Wang, Occluded person re-identification, in: 2018 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2018, pp. 1–6.
[50] Jiaxuan Zhuo, Jianhuang Lai, and Peijia Chen. A novel teacher-student learning framework for occluded person re-identification. ArXiv:1907.03253, 2019..